Action Spécifique Geste et Action : Compte-rendu de la réunion du 15/06/2007

Dans le cadre de l'action spécifique Geste et Action du GDR ISIS, nous avons organisé une deuxième réunion le 15 juin à l'ENST sur les thèmes de l'acquisition/suivi de geste, et la reconnaissance d'actions et d'expressions. Elle a donné lieu à sept exposés. La journée a été suivie par plus de 35 participants et a confirmé l'intérêt suscité par cette action spéciifique.
Les présentations sont en ligne à : http://imtr.ircam.fr/index.php/AS_Geste_15_juin_2007

Organisateurs : Frederic Bevilacqua, Catherine Achard, Patrick Horain

Programme

Acquisition des gestes, multimodalité (matin)

1**) Stochastic search for monocular articulated body pose estimation**
Benedicte Bascle, France Telecom R&D

2) **Estimation de la pose 3D d'un modèle déformable générique à partir de séquences d'images monoculaires**
Martin de La Gorce , Ecole Centrale Paris

3) **Movimento: vers l"Extreme Mocap"**
Nicolas Gros et Loic Lefort, REALVIZ

4**) Interaction multimodale : Faites un geste, parlez lui!**
Jean Emmanuel Viallet France Telecom R&D

Reconnaissance de gestes et d'actions, Visages (après-midi)

5) **Reconstruction et Animation de Visage**
Charlotte Ghys Ecole Centrale Paris, France Telecom R&D

6**) Reconnaissance d'actions de la vie quotidienne**
Catherine Achard et Arash Mokhber Universite Pierre et Marie Curie (Paris 6)

7) **View Independent Action Recognition**
Daniel Weinland INRIA Rhone-Alpes

Résumés:

1) **Stochastic search for monocular articulated body pose estimation**
Benedicte Bascle, , France Telecom R&D

This talk presents an algorithm for automatic probabilistic inference of human upper body pose and appearance in a scene. The approach consists in looking at pairs of randomly selected (and distant) images from the sequence. For each image pair $(I_t,I_{t'})$, a stochastic search is made to find pairs of likely upper body limb positions. The search uses a MCMC method with a (quasi-) independent proposal mechanism based on limb detection and designed to maximise state space coverage. The state posterior probability combines image-based cues and a measure of similarity of body appearance (color, texture, ...) between the

two images. Because noise and outliers in distant images are unlikely to be coherent, false positives given by limb detection in each image do not usually have similar appearance to false positives in the other image in the pair. Thus they are discarded and only real positives have good likelihood. These give both upper body pose and limbs appearances models. A strength of the approach is that it can be applied to complex sequences where people appearance is unknown and where motion is at times too fast or unexpected for tracking approaches to perform well. It can be applied to people with different builds and types of clothing. Because it does not rely on pre-learning possible models of appearance (and associated edge maps for instance), it is less likely to fail when atypical appearances are seen. The obtained appearance and body poses can be used to initialize an adapted tracking algorithm to find the body pose during the whole sequence.

2) Estimation de la pose 3D d'un modèle déformable générique à partir de séquences d'images monoculaires

Martin de La Gorce , Ecole Centrale Paris

Une nouvelle formulation variationelle pour déterminer la pose 3D d'un modèle déformable générique à partir de séquences d'images monoculaires est présentée. Le problème est formulé comme un problème inverse, où la fonction minimisée est basée sur une représentation triangulée de la surface et permet de modéliser explicitement les auto-occlusions, la présence de texture et les ombrages pour des sources de lumière qui peuvent varier au cours du temps. Les paramètres correspondants font partie du modèle et sont ré-estimés à caque nouvelle image par descente de gradient sur la fonction objective. Nous présenterons le calcul exact du gradient de la fonction objective dans le domaine continu, en détaillant plus particulièrement les termes qui sont dus aux déplacements des auto-occlusions et qui sont généralement négligés dans la littérature. Cela nous amènera à introduire des nouvelles forces le long des occlusions qui améliorent la performance de la méthode lorsque l'objet suivit est peu texturé. Pour valider la méthode proposée nous l'appliquons au problème de suivit de la main à partir de séquence d'images monoculaires. Les résultats obtenus semblent démontrer la pertinence de notre méthode

3) **Movimento: vers l'"Extreme Mocap"**

Nicolas Gros et Loic Lefort, REALVIZ

Movimento est le nouveau système de capture de mouvement à partir de vidéos développé par REALVIZ. La capture de mouvement optique offre certaines limitations: - marqueurs spécifiques - environnement contraint (éclairage) - taille de champ limité - coût - difficulté de mise en place - nécessité d'un expert La maturité de la vision par ordinateur offre désormais la possibilité d'opérer avec beaucoup plus de flexibilité et de légèreté en s'affranchissant d'une partie de ces contraintes. Movimento est né de ce constat et du savoir-faire de REALVIZ en terme de traitement d'images et de reconstruction 2D/3D. Cette solution permet de réaliser une capture de mouvement de qualité à partir de vidéos standard, en utilisant des caméras potentiellement mobiles, voire à focale variable, et un éclairage non contrôlé. Cette flexibilité ouvre la voie vers des applications nouvelles dites d'"extreme mocap". Nous présenterons Movimento ainsi que quelques exemples d'applications impossibles à réaliser sans cette souplesse: - on-set mocap - mocap en champ réduit - sports en extérieur -...

4) **Interaction multimodale : Faites un geste, parlez lui!**

Jean Emmanuel Viallet France Telecom R&D

La vision par ordinateur, la reconnaissance de la parole permettent une interaction naturelle, sans autre périphérique que son corps, avec des écrans de notre taille. On a toujours avec soi et sa voix et ses mains. Ce sont des interfaces que l'on maîtrise depuis l'enfance, que l'on n'a pas besoin d'apprendre. Ils sont toujours disponibles, non encombrant ; il est inutile de les partager car chacun possède les siens. Rien qu'avec les mains, l'utilisateur contrôle son ordinateur, en retrouvant les fonctionnalités d'une souris : le pointage et la sélection. La reconnaissance de parole permet une interaction multimodale, de joindre le geste à la parole. Il s'agit alors de synchroniser les deux modalités, puis de les fusionner en prenant en compte éventuellement le contexte. Les performances de dispositif d'interaction gestuelle et oro-gestuelle sans contact sont évaluées selon la norme ISO 9241-9:2000. Le système MOWGLI (Multimodal Oral With Gesture Large display Interface) sera illustré par plusieurs vidéos où, un ou deux utilisateurs interagissent avec le système sans autre périphérique que leur propre corps, par exemple pour jouer aux échecs.

5) Reconstruction et Animation de Visage
Charlotte Ghys Ecole Centrale Paris, France Telecom R&D

Je présenterai mon travail sur le mimétisme 3D d'expressions faciales à partir d'une séquence monoculaire. J'introduirai tout d'abord le maillage sémantique que nous avons construit, à partir d'une base de donnée de reconstructions 3D de visages. Ce modèle, basé sur le standard MPEG-4, est animé par des points de contrôle et sa surface est déformée en utilisant les Radial Basis Functions. Pour déterminer la position 2D de ces points de contrôle dans la séquence, nous utilisons l'algorithme Adaboost en cascade, contraint par une modélisation dynamique des expressions. Parmi les candidats fournis par Adaboost, on extrait une combinaison optimale par programmation linéaire combinatoire, qui force un positionnement anthropométrique des points d'intérêt. Leur mouvements peuvent alors être reproduits sur n'importe quel autre visage.

6) **Reconnaissance d'actions de la vie quotidienne**
Catherine Achard et Arash Mokhber Universite Pierre et Marie Curie

Nous proposons lors de cette présentation de reconnaître des actions usuelles comme « marche », « s'assoit sur une chaise », « saute », « se penche » ou « s'accroupit » en utilisant une seule caméra sans calibrage. L'invariance au point de vue est gérée grâce à plusieurs vues de la même action, ce qui évite de faire appel à des informations 3D. Une détection de mouvement est tout d'abord réalisée sur chaque image, conduisant à une série d'images binaires. Deux approches ont ensuite été envisagées, chacune présentant ses avantages et défauts : - Une modélisation globale des séquences où tous les points binaires formant dans l'espace 3D (x,y,t) un volume sont caractérisés par leurs moments géométriques 3D. Ces moments sont normalisés afin d'être invariants en translation et en changement d'échelle. La reconnaissance d'actions est ensuite effectuée en calculant la distance de Mahalanobis entre le vecteur de caractéristiques de l'action à reconnaître et ceux de la base de référence. - Une normalisation par des caractéristiques semi-globales (extraites sur plusieurs images de la séquence). Les moments géométriques 3D calculés à partir du volume constitué par tous les points détectés en mouvement dans la fenêtre étudiée sont estimés. Ces caractéristiques amènent, pour chaque action, à une chaîne temporelle, qui constitue l'entrée d'un système de reconnaissance utilisant les chaînes de Markov cachés. Des résultats de reconnaissance d'actions seront présentés pour les deux méthodes sur une base de 1662 séquences réalisées par plusieurs personnes et réparties en huit classes. Des taux de reconnaissance de l'ordre de 94% sont obtenus.

7) **View Independent Action Recognition**
Daniel Weinland INRIA Rhone-Alpes

I will present my work on view-independent action recognition: - "Motion History Volumes" (MHV) are a free-viewpoint representation for human actions in the case of multiple calibrated, and background-subtracted, video cameras. We present algorithms for segmentation and classification of different actions performed by different people in a variety of viewpoints. Results indicate that MHVs can be used to learn and recognize basic human action classes, independently of gender, body size and viewpoint. - I will present recent work that uses four dimensional action models, learned from multiple views, to recognize actions from a single or few unknown viewpoints.

.